SkinCON: Towards consensus for the uncertainty of skin cancer sub-typing through distribution regularized adaptive predictive sets (DRAPS)

Zhihang Ren[1,†], Yunqi Li[1,*], Xinyu Li[1,*], Xinrong Xie[2], Erik P. Duhaime[2], Kathy Fang[3], Tapabrata Chakraborty[4,5], Yunhui Guo[6], Stella X. Yu[7], and David Whitney[1]

[1] University of California, Berkeley, CA 94720, USA
[2] Centaur Labs, Boston MA, USA
[3] Golden State Dermatology, Albany CA, USA
[4] The Alan Turing Institute and University College London, London, UK
[5] University of Oxford, Oxford, United Kingdom
[6] University of Texas at Dallas, Richardson, TX 75080, USA
[7] University of Michigan, Ann Arbor, MI 48109, USA

Abstract. Deep learning has been widely utilized in medical diagnosis. Convolutional neural networks and transformers can achieve high predictive accuracy, which can be on par with or even exceed human performance. However, uncertainty quantification remains an unresolved issue, impeding the deployment of deep learning models in practical settings. Conformal analysis can, in principle, estimate the uncertainty of each diagnostic prediction, but doing so effectively requires extensive human annotations to characterize the underlying empirical distributions. This has been challenging in the past because instance-level class distribution data has been unavailable: Collecting massive ground truth labels is already challenging, and obtaining the class distribution of each instance is even more difficult. Here, we provide a large skin cancer instance-level class distribution dataset, SkinCON, that contains $25,331$ skin cancer images from the ISIC 2019 challenge dataset. SkinCON is built upon over $937,167$ diagnostic judgments from $10,509$ participants. Using SkinCON, we propose the distribution regularized adaptive predictive sets (DRAPS) method for skin cancer diagnosis. We also provide a new evaluation metric based on SkinCON. Experiment results show the quality of our proposed DRAPS method and the uncertainty variation with respect to patient age and sex from health equity and fairness perspective. The dataset and code are available at https://skincon.github.io.

Keywords: Conformal prediction · Skin cancer dataset · Diagnostic trial · Uncertainty Quantification · Health Equity and Fairness

1 Introduction

Deep learning and computer vision approaches are increasingly employed in medical practice to assist clinicians in their decisions. Recent artificial intelligence

[*] Equal contribution; [†] Corresponding author: peter.zhren@berkeley.edu

diagnostic algorithms can even rival human levels of performance [9, 14], with state-of-the-art skin cancer diagnostic models achieving over 90% accuracy [2, 7]. However, it is hard to deploy those superior AI models into realistic clinical scenarios; mistrust is a major barrier to clinical implementation of deep learning predictions [13, 21].

Imagine you are a doctor who is making an important diagnostic decision, trying to determine what type of skin cancer a particular lesion sample is. You are provided with a class label produced from several state-of-the-art computer vision models. With potential biases and unexplainable errors [23], would you choose to trust those state-of-the-art computer vision models? Currently, a maximum likelihood diagnosis (which most classifiers adopt)— even with the known overall performance of the classifier and an accompanying probability—may not be the essential information for you. To make an accurate diagnosis, you may want to consider all potential disease types, and you might want to take into account the possibility of an especially harmful (mis)diagnosis or lesion type. Thus, in addition to an estimate of the most likely outcome, you would like the classifier to also offer you *actionable uncertainty quantification.* This can be a set of predictions that provably covers the true diagnosis with a high probability (e.g., 90%), and it is called a prediction set (see Figure 1).

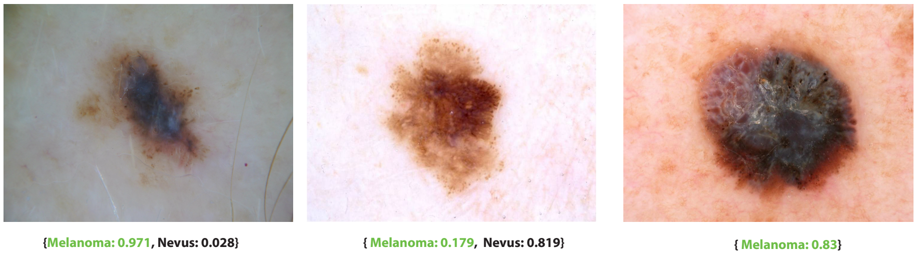
{Melanoma: 0.971, Nevus: 0.028}
{ Melanoma: 0.179, Nevus: 0.819}
{ Melanoma: 0.83}

Fig. 1. Prediction set examples on ISIC 2019 challenge dataset. We show three examples of the class Melanoma and the 90% prediction sets generated by DRAPS.

Conformal analysis [24] is a rigorous statistical method to generate such a prediction set, i.e., the conformal prediction set. Formally, imagine we have n data samples $\{(X_i, Y_i)\}_{i=1}^{n}$ with features $X_i \in \mathbb{R}^p$ and a discrete label $Y_i \in \mathcal{Y} = \{1, 2, ..., C\}$. The samples are drawn exchangeably (e.g., i.i.d., although exchangeability alone is sufficient) from some unknown distribution P_{XY}. Given such data and a desired coverage level $1-\alpha \in (0, 1)$, we seek to construct a prediction set $\hat{C}_{n,\alpha} \subseteq \mathcal{Y}$ for the unseen label of a new data point (X_{n+1}, Y_{n+1}), also drawn exchangeably from P_{XY}, achieving marginal coverage; that is, obeying

$$\mathbb{P}[Y_{n+1} \in \hat{C}_{n,\alpha}(X_{n+1})] \geq 1-\alpha \tag{1}$$

The probability above is taken over all n + 1 data points, and we ask that Inequality 1 holds for any fixed α, n, and P_{XY}. Traditional conformal predictors [18, 11, 16, 19, 1, 15, 3] can modify any black-box classifier to output predictive sets that are rigorously guaranteed to satisfy the desired coverage property shown in Inequality 1. These traditional conformal predictors generally need a calibration dataset that can help the algorithm establish thresholds based on different desired coverage levels. Then the conformal prediction set can be generated by comparing the logits from the networks to the established threshold.

This kind of formal quantification of uncertainty is the need of the hour for high-stakes biomedical applications of AI, like computational cancer sub-typing: a guarantee of predictive robustness would provide trustable interpretation for clinicians to adopt AI-based models as valid clinical decision support systems. A challenge here is that skin lesions are highly heterogeneous, and hence out-of-distribution samples are often the norm rather than being rare outliers. Moreover, skin cancer diagnosis can suffer from a lack of consensus among clinicians. Because conformal analysis assumes that the class labels provided are trustable, what is needed is something more flexible than standard conformal prediction. We therefore propose a 'conformal inspired' approach, which starts off with a similar formulation, but then modifies it to match a distribution of labels from multiple experts.

To achieve this, the present work makes two significant contributions. First, we collect and curate a new multi-label skin cancer dataset SkinCon, which starts with the benchmark ISIC 2019 repository but presents it to a group of proficient annotators (e.g., clinicians) to label each sample, thus, creating a label distribution for consensus building. Second, on the methodological end, we start with the conformal formulation but regularize it to match the label distribution, while still maintaining state-of-the-art coverage and conformal set size. We name this novel method DRAPS (distribution regularized adaptive prediction sets), which, to our knowledge, is the first of its kind for consensus building in cancer sub-typing. In addition to applications in skin cancer, DRAPS is model agnostic and may be used for any such cancer sub-typing scenario. Finally, in this study, with the instance-level consensus that SkinCON provided, we propose a new evaluation metric — a "hit rate" to evaluate the quality of the prediction set.

2 SkinCON

SkinCON contains 25, 330 skin cancer images from the ISIC 2019 challenge dataset [22, 4, 5] and is built upon over 937, 167 diagnostic trials from 10, 509 participants. The skin lesion diagnoses were collected through DiagnosUs, an app developed by Centaur Labs, a US medical Artificial Intelligence (AI) company based in Boston, MA. SkinCON dataset contains every skin cancer image information about the filename reference to the ISIC 2019 challenge dataset, the number of qualified reads, the correct label, the majority label, difficulty, agreement, and the corresponding number of qualified reads for 8 skin cancer types. The 8 skin cancer types include: actinic keratosis, basal cell carcinoma, benign keratosis, dermatofibroma, melanoma, nevus, squamous cell carcinoma, and vascular lesion. We noticed there is another SkinCON Dataset [6]. However, they focus on fine-grained analysis, while ours aims on uncertainty estimation.

SkinCON directly provides us with the instance-level empirical response distributions of lesion images. Collecting a massive quantity of human responses reveals resulting distributions that naturally depict the inherent uncertainty of each skin cancer image. Figure 2 shows the diverse response distributions of sample skin lesion images from different categories. Even from the same category, the response distribution can vary a lot. However, with noisy responses, the consensus (accuracy after popularity voting) is 61.86% (v.s., chance level of 12.5%). It means that our data captures important information about the response distribution.

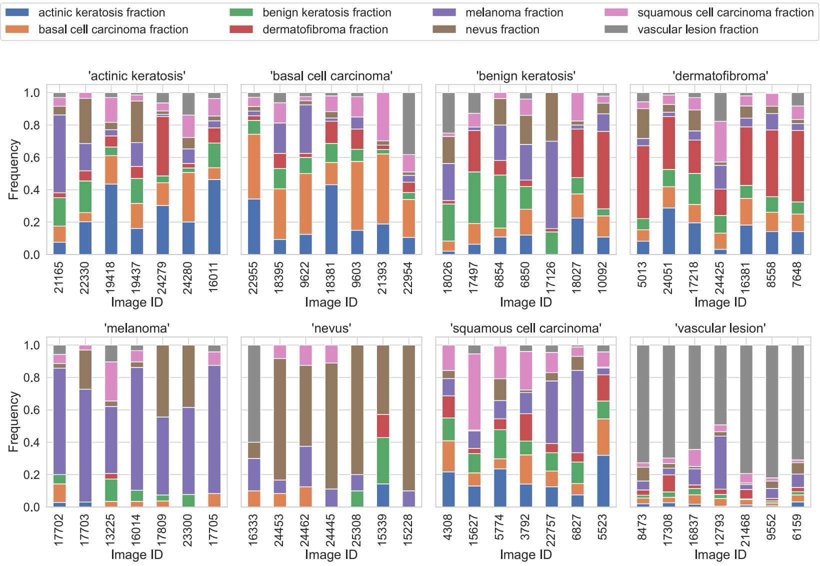

Fig. 2. Response distributions on ISIC 2019 challenge dataset. We plot the response frequency of sample skin cancer images. We argue that the empirical response distribution may reveal the inherent skin cancer property.

2.1 Health Fairness Study

The ISIC 2019 challenge dataset also provides us with the demographic data of patients. Using patients' demographic information, we group the skin cancer images with regard to gender and age, respectively. We then check the diagnostic accuracy for each category of lesion. Results are shown in Figure 4 and Figure 3.

We find there are significant differences between patient gender in the cases of benign keratosis ($p < 0.001$), nevus ($p < 0.001$), vascular lesion ($p < 0.05$), and overall skin cancer categories ($p < 0.01$). The significance test was conducted by comparing the empirical mean difference value with the null distribution of differences, corrected for multiple comparisons.

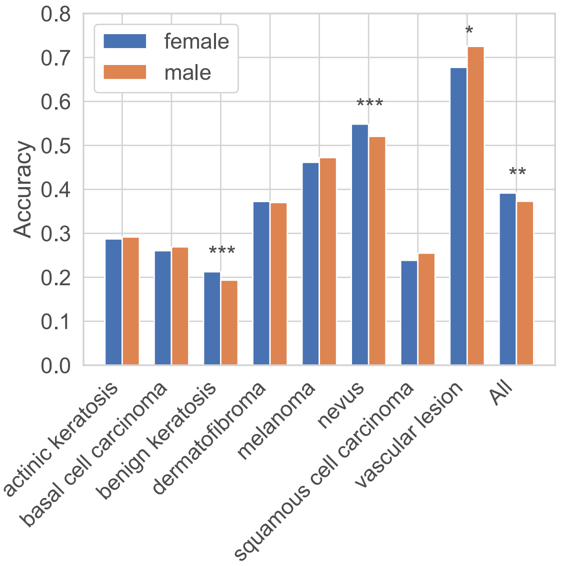

Fig. 3. Diagnostic biases as a function of patient gender in ISIC 2019 challenge dataset. Significant differences were found as a function of patient gender for benign keratosis, nevus, vascular lesion, and overall skin cancer categories. $*p < 0.05$, $**p < 0.01$, $***p < 0.001$

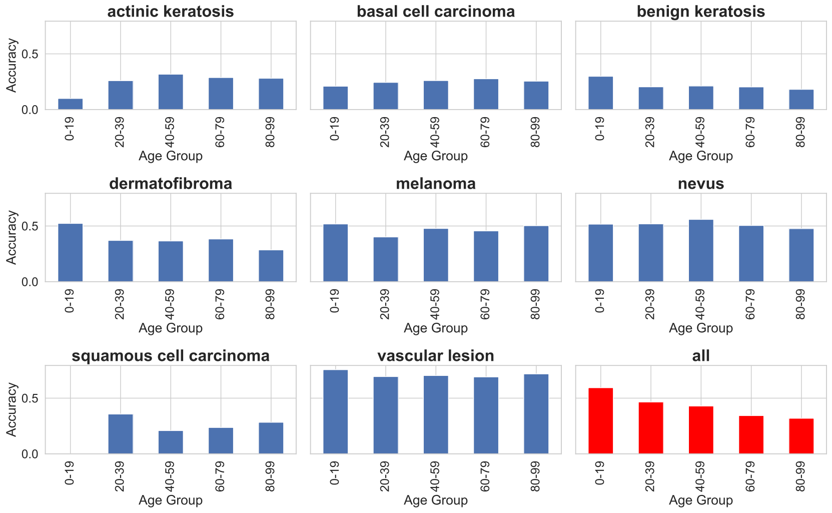

Fig. 4. Diagnostic biases as a function of patient age in ISIC 2019 challenge dataset.

For different age groups, we find the biases are different for different skin cancer categories. In particular, we find overall diagnostic accuracy decreases as patient age increases. The reasons for this are unknown but could involve many different factors and possible confounds. In any case, taking into account these differences may be important.

2.2 Data Annotation Procedure

The participants were mostly composed of medical students, with some medical residents. Individual subject information such as age or sex is not known. All participants have normal or corrected-to-normal vision. Users receive earnings from a predefined money pool (around 50 USD) for each task they complete.

After downloading the DiagnosUs app and giving consent to have Centaur Labs use the data they provide through app usage, users can choose between different tasks. For the dermatological classification task that was investigated in this study, users first completed a training session of 10 trials with 10 separate stimuli. This training explained the procedure of the task and prepared users for the actual classification task, which was identical to the training. In each trial, a random skin lesion image was selected and presented to the participant. Below the image, they were prompted to choose one of the eight possible skin cancer types. Feedback was provided after every trial to inform users if their response was correct or incorrect. Afterward, users voluntarily moved on to the next trial at their own pace. Users were told they could end the task at any time.

Data monitoring, storage, and safety procedures were carried out in accordance with university-approved IRB protocols.

3 Distribution Regularized Adaptive Prediction Sets

SkinCON allows us to directly learn the instance-level distribution end-to-end via deep learning models. In this study, we propose the distribution regularized adaptive prediction sets, DRAPS. For any deep learning diagnostic backbones, we train it via general cross-entropy loss $L_{Entropy}$ with additional Kullback–Leibler (KL) divergence loss L_{KL}, $L = L_{Entropy} + \lambda L_{KL}$. The KL divergence measures how different two distributions are. The larger the KL divergence loss, the more different the two distributions are. During training, the model parameters are optimized to minimize the KL divergence loss that forces the learned distribution to be similar to the empirical distribution. With the help of KL-divergence loss, classifiers can match the empirical distribution of training instances.

We directly calibrate our proposed method on the training dataset. Given the softmax logits $s \in [0,1]^{n \times K}$ and ground truth labels for each of n examples in the training set $y \in \{0, 1, ..., K\}^n$ with K possible classes. We can find the softmax logit of each example's ground truth label E_i, i.e. $s_{i,j}$ where $j = y_i$. Then, the $(\lfloor \alpha * n \rfloor - 1)$-th smallest value in $\{E_i\}_{i=1}^{n}$ is the threshold τ of our proposed method. Algorithm 1 summarises this module.

Algorithm 1 Distribution Regularized Adaptive Prediction Calibration

```
    Input: $\alpha$; $s \in [0,1]^{n \times K}$, $y \in \{0, 1, ..., K\}^n$ respectively to the scores, and ground truth labels for each of n examples in the training set.
  procedure DRAPC($\alpha, s, y$)
      for $i \in \{1, ..., n\}$ do
          $E_i \leftarrow s_{i,j}$ such that $j = y_i$
      $\tau \leftarrow$ the $(\lfloor \alpha * n \rfloor - 1)$-th smallest value in $\{E_i\}_{i=1}^n$
      return $\tau$
    Output: Threshold $\tau$
```

After the threshold τ is obtained, we can generate the prediction set C for each sample accordingly. For each test sample's softmax logits s_i, we compare it with the obtained threshold τ. The corresponding label will be appended to the prediction set C if the softmax logit is greater or equal to the threshold, i.e., satisfying $s_i \geq \tau$. Algorithm 2 summarises this module.

Algorithm 2 Distribution Regularized Adaptive Prediction Sets

```
    Input: $\alpha$, the scores $s$ for a test-time example, threshold $\tau$ from Algorithm 1.
  procedure DRAPS($\alpha, s, \tau$)
      $C \leftarrow \{\}$
      for $i \in \{1, ..., K\}$ do
          if $s_i \geq \tau$ then
              $C.append(s_i)$
          return $C$
    Output: The 1 - $\alpha$ confidence set, $C$
```

3.1 Hit Rate

SkinCON contains instance-level response distribution, therefore, it can provide an ordering of skin cancer categories for a specific lesion. Here, we propose to utilize Hit Rate (HR) that utilizes this ordering to evaluate the quality of certain prediction sets. Given the prediction set C with size k and the top k responses of this skin cancer image R_k, Hit Rate (HR) $= len(C \cap R_k)/k$. Intuitively, the Hit Rate will be maximized when the prediction set with length k exactly contains the top k responses.

4 Experiments and Results

We compare three methods: a naive baseline (Naive), regularized adaptive prediction sets (RAPS) [1], and our proposed DRAPS. The naive baseline is basically directly utilizing the standard cross-entropy loss to train the diagnostic task without matching the empirical response distribution. The latter calibration and prediction set generation processes are the same as our proposed DRAPS.

We adopt multiple image classification backbones for experiments, including variants of ResNet [10], ResNeXt [25], VGG [20], ShuffleNet [26], and DenseNet [12].

4.1 Implementation Details

All experiments are implemented through PyTorch platform [17] and trained on a single NVIDIA GeForce RTX 2080 Ti. The learning rate is set to 0.001 for all model training with a stochastic gradient descent (SGD) optimizer. $\lambda = 0.1$ for the KL-divergence loss. The training epoch is 30 for all experiments, and the batch size is 8. The training and testing data are split with a ratio of $9 : 1$ for all classes.

4.2 Prediction Set Results

We test the proposed distribution regularized adaptive prediction sets (DRAPS) at $\alpha = 0.1$. Results are shown in Table 1. The top-1 and top-5 accuracies indicate the performance of those backbone models. Our DRAPS can reach the desired coverage level while utilizing the smallest average prediction set size. It is noted that modern models like vision transformers (ViTs) [8] have been utilized in medical image recognition tasks. We expect the performance gain would be less with better baseline models.

Model	Accuracy		Coverage			Size		
	Top-1	Top-5	Naive	RAPS	Ours	Naive	RAPS	Ours
ResNet18	91.62	99.94	0.928	0.964	0.935	1.278	1.183	**1.054**
ResNet50	92.07	100.0	0.935	0.967	0.941	1.165	1.192	**1.104**
ResNet101	93.02	100.0	0.947	0.961	0.931	1.367	1.163	**1.047**
ResNet152	91.50	99.94	0.934	0.961	0.936	1.241	1.167	**1.021**
ResNeXt101	92.77	99.87	0.938	0.972	0.941	1.050	1.179	**1.006**
VGG16	91.52	99.87	0.927	0.961	0.924	1.056	1.213	**1.008**
ShuffeNet	89.85	99.56	0.923	0.968	0.926	1.241	1.414	**1.154**
DenseNet161	92.83	99.94	0.941	0.971	0.936	1.141	1.146	**1.034**

Table 1. Results on ISIC 2019 Challenge Dataset with $\alpha = 0.1$ We report coverage and size of naive, RAPS, and our proposed method for eight different image classifiers.

4.3 Hit Rate as a New Evaluation Metric

Based on SkinCON, we propose Hit Rate as a new evaluation metric to quantify the quality of the prediction set. Results from different methods are shown in Table 2. DRAPS achieves the best hit rate, benefitting from learning the empirical response distribution.

5 Conclusion

In this paper, we release the SkinCON, a large skin cancer instance-level class distribution dataset for decision consensus building. We also propose the distribution regularized adaptive prediction sets (DRAPS) to increase the predictive

Model	Hit Rate		
	Naive	RAPS	Ours
ResNet18	0.852	0.902	**0.905**
ResNet50	0.884	0.901	**0.902**
ResNet101	0.849	0.900	**0.905**
ResNet152	0.866	0.908	**0.912**
ResNeXt101	0.913	0.915	**0.920**
VGG16	0.904	**0.905**	**0.905**
ShuffeNet	0.863	0.878	**0.884**
DenseNet161	0.892	0.906	**0.910**

Table 2. Hit Rates on ISIC 2019 Challenge Dataset with $\alpha = 0.1$ We report the hit rates of naive, RAPS, and our proposed method for eight different image classifiers.

robustness. Finally, we propose the hit rate as a new evaluation metric to quantify the quality of the prediction sets. We visualize the uncertainty variation with respect to patient age and sex from a health equity and fairness perspective.

Acknowledgments. This study was funded by the National Institutes of Health (NIH) under grant R01CA236793.

Disclosure of Interests. XX and ED of Centaur Labs developed and have a financial interest in the DiagnosUs app used to collect the data. KF has provided services for Golden State Dermatology. The other authors have no competing interests.

References

1. Angelopoulos, A.N., Bates, S., Jordan, M., Malik, J.: Uncertainty sets for image classifiers using conformal prediction. In: International Conference on Learning Representations (2020)
2. Benčević, M., Galić, I., Habijan, M., Babin, D.: Training on polar image transformations improves biomedical image segmentation. IEEE Access **9**, 133365–133375 (2021)
3. Cauchois, M., Gupta, S., Duchi, J.C.: Knowing what you know: valid and validated confidence sets in multiclass and multilabel prediction. The Journal of Machine Learning Research **22**(1), 3681–3722 (2021)
4. Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). pp. 168–172. IEEE (2018)
5. Combalia, M., Codella, N.C., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Carrera, C., Barreiro, A., Halpern, A.C., Puig, S., et al.: Bcn20000: Dermoscopic lesions in the wild. arXiv preprint arXiv:1908.02288 (2019)

6. Daneshjou, R., Yuksekgonul, M., Cai, Z.R., Novoa, R., Zou, J.Y.: Skincon: A skin disease dataset densely annotated by domain experts for fine-grained debugging and analysis. Advances in Neural Information Processing Systems **35**, 18157–18167 (2022)
7. Datta, S.K., Shaikh, M.A., Srihari, S.N., Gao, M.: Soft attention improves skin cancer classification performance. In: International Workshop on Interpretability of Machine Intelligence in Medical Image Computing. pp. 13–23 (2021)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)
9. Haenssle, H.A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Hassen, A.B.H., Thomas, L., Enk, A., et al.: Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Annals of oncology **29**(8), 1836–1842 (2018)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
11. Hechtlinger, Y., Póczos, B., Wasserman, L.: Cautious deep learning. arXiv preprint arXiv:1805.09460 (2018)
12. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
13. Linegang, M.P., Stoner, H.A., Patterson, M.J., Seppelt, B.D., Hoffman, J.D., Crittendon, Z.B., Lee, J.D.: Human-automation collaboration in dynamic mission planning: A challenge requiring an ecological approach. In: Proceedings of the human factors and ergonomics society annual meeting. vol. 50, pp. 2482–2486. SAGE Publications Sage CA: Los Angeles, CA (2006)
14. Mar, V., Soyer, H.: Artificial intelligence for melanoma diagnosis: how can we deliver on the promise? Annals of Oncology **29**(8), 1625–1628 (2018)
15. Messoudi, S., Rousseau, S., Destercke, S.: Deep conformal prediction for robust models. In: 18th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU 2020). pp. 528–540 (2020)
16. Park, S., Bastani, O., Matni, N., Lee, I.: Pac confidence sets for deep neural networks via calibrated prediction. In: International Conference on Learning Representations (2019)
17. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch (2017)
18. Platt, J., et al.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers **10**(3), 61–74 (1999)
19. Romano, Y., Sesia, M., Candes, E.: Classification with valid and adaptive coverage. Advances in Neural Information Processing Systems **33**, 3581–3591 (2020)
20. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
21. Stubbs, K., Hinds, P.J., Wettergreen, D.: Autonomy and common ground in human-robot interaction: A field study. IEEE Intelligent Systems **22**(2), 42–50 (2007)

22. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**(1), 1–9 (2018)
23. Varoquaux, G., Cheplygina, V.: Machine learning for medical imaging: methodological failures and recommendations for the future. NPJ digital medicine **5**(1), 48 (2022)
24. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic learning in a random world, vol. 29. Springer (2005)
25. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1492–1500 (2017)
26. Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6848–6856 (2018)